

Text S1. Expected relative contribution of BER to simple and complex conversion tracts.

To distinguish between the BER and MMR models, we compared the strength of gBGC in simple and complex conversions tracts (a complex conversion tracts is a tract involving conversion events from both parental haplotypes, whereas simple tracts involve only one donor haplotype). Indeed, one clear difference between BER and MMR is the length of the region affected by the repair: whereas MMR involves DNA re-synthesis over hundreds of base pairs (i.e. about the size of conversion tracts) (Winzeler et al. 1998; Holmes and Clark 1990), BER leads only to short-patch repair (1-13bp) (Evans and Alani 2000). Given that the median distance between contiguous markers is 78 bp, if some SNP conversion events are driven by BER, then the conversion of these SNPs should occur independently of the conversion of neighboring SNPs. Thus, if BER is active during recombination, it is expected to lead frequently to complex conversion tracts.

To justify this assertion more formally, let us consider a simple model, in which conversion tracts can be affected either by MMR only (probability p), or by both MMR and BER (probability $1-p$), with BER acting independently on n SNPs in the tract ($n>0$). When both MMR and BER are acting on the tract, then for each SNP repaired by BER, the probability to be repaired using the same template strand as MMR is 0.5. Thus, when BER is active, the expected proportions of simple and complex conversion tracts are respectively (0.5^n) and $(1 - 0.5^n)$. BER is probably not the unique cause of complex conversion tracts. Let us assume that in absence of BER, MMR can lead to complex tracts with probability q . We can now express the proportion of complex (f_{cpx}) and simple (f_{smp}) tracts:

$$\begin{cases} f_{cpx} = p (1 - 0.5^n) + (1 - p) q \\ f_{smp} = p 0.5^n + (1 - p)(1 - q) \end{cases} \quad (1)$$

Thus, the proportion of conversion tracts in which BER was active, among complex tracts (BER_{cpx}) and among simple tracts (BER_{smp}), can be expressed as follows:

$$\begin{cases} BER_{cpx} = p (1 - 0.5^n) / f_{cpx} \\ BER_{smp} = p 0.5^n / f_{smp} \end{cases} \quad (2)$$

In the Mancera dataset [23], the proportion of complex events is $f_{cpx} = 8.8\%$. Thus, the ratio f_{smp}/f_{cpx} is about 10. Hence, we can deduce that:

$$\frac{BER_{cpx}}{BER_{smp}} \approx 10 (2^n - 1) \quad (3)$$

Given that $n \geq 1$, the value of this ratio is at least 10. Thus, under the assumption that a fraction of SNP conversions results from the activity of BER, then the proportion of tracts in which BER was involved should be at least 10 times higher among complex tracts than among simple tracts. Hence, if gBGC was due to BER, then the signal of gBGC should be much stronger among complex conversion tracts as compared to simple ones.

Text S2. Conversion bias towards GC-rich haplotypes (defined with a more stringent threshold).

In the Figure 2, we considered all conversion events for which there was a difference in GC-content between the two haplotypes, even if this difference was weak. In order to test if this categorization does not introduce any bias in our study, we repeated the same analysis using only conversion events for which there is a strong difference in GC-content between the two haplotypes. To measure this difference, we introduced the ΔGC_h parameter, which is defined as follow:

$$\Delta GC_h = \frac{|GC_Y - GC_S|}{n} \quad (4)$$

Where n is the number of AT/GC SNP sites located in the conversion tract, and GC_Y and GC_S , are the numbers of GC-alleles at those sites respectively for the YJM789 and S96 strains. In this section, we only used events for which $\Delta GC_h > 1/3$ (N=1,801 recombination events). Using this threshold, the intensity of the bias (b) remains the same as in previous analyses (compare Figure S2 with Figure 2) for all three categories (All, CO and NCO). This shows that our results are not affected by the threshold used to define GC-rich/AT-rich haplotypes.

Text S3. Model of gBGC driven by GC-biased MMR-repair: expected relative abundance and lengths of conversion tracts with GC_f or AT_f haplotype.

Recombination is initiated by the formation of a DSB on one chromatid, followed by 5' to 3' resection, which produces two single-stranded 3' overhangs. The DSB is subsequently repaired using an intact template (sister chromatid or homolog) (Figure 1). This DSB repair involves the formation of heteroduplex DNA, with one strand coming from the template chromatid and the other corresponding to one single-stranded 3' overhang of the broken chromatid. If the template chromatid is the homolog, then the two haplotypes in the heteroduplex may be different. The two corresponding haplotypes will hereafter be referred to as the "template haplotype" and the "broken haplotype". Mismatches in the heteroduplex can be repaired and, depending to the direction of repair, this process leads either to conversion of the broken haplotype by the template haplotype, or to restoration of Mendelian segregation.

Let us consider a heteroduplex with AT_f/GC_f polymorphism, i.e. in one haplotype the alleles located at the extremities of the heteroduplex are A or T, whereas in the other they are G or C (Figure 4). Let us note p_s the probability of conversion when the template haplotype is GC_f and p_w the probability of conversion when the template haplotype is AT_f. The repair of mismatches by MMR is nick-directed. By definition, one nick is always present on the strand corresponding to the broken haplotype. But the resolution of recombination intermediates may also involve the formation of nicks on the other strand. According to our model, in that situation, the direction of repair is biased towards the GC_f haplotype. Hence, $p_s > p_w$.

Many of the recombination pathways involve the invasion of the template chromatid by both single-stranded 3' overhangs of the broken chromatid, and hence involve the formation of two heteroduplex DNA (Figure 1). The extent of the detected conversion tract (and hence the nature of the SNPs at its boundaries) depends on the direction of repair (conversion or restoration) on both heteroduplex. We assume that for a given haplotype, the nature of the alleles (GC or AT) at the polymorphic sites located in one heteroduplex is independent of the alleles present in the other heteroduplex (i.e. we assume that there is no correlation between the nature of neighboring alleles on a same haplotype). Under this assumption, the direction of repair (restoration or conversion) in one heteroduplex is independent of the direction of repair in the other one.

In Figure 4, we present the different conversion tracts that can be obtained, depending on whether each of the two heteroduplex is subject to conversion or restoration. When both heteroduplex are subject to restoration, no conversion tract can be observed. The other cases can lead to 5 different configurations, with relatively long (length = L_2) or short (length = L_1 , with $L_2 > L_1$) conversion tracts. In Figure 4, we present in detail the different possible outcomes of mismatch repair by MMR for one heteroduplex (heteroduplex II). Here we focus on cases where all mismatches in the heteroduplex are repaired in the same direction (i.e. we focus on cases that lead to simple conversion tracts, which represent the vast majority of recombination events; see Main Text). For the sake of simplicity, we consider that in all cases, the second heteroduplex (heteroduplex I) is subject to conversion with probability x or restoration with probability $(1 - x)$. We assume that there is no initiation bias, i.e. the frequency of cases where the broken haplotype is AT_f (case A in Figure 4) is equal to the frequency of cases

where the broken haplotype is GC_f (case B). Under those assumptions, it is possible to compute the relative abundance of each of the five configurations of detectable conversion tracts (Table S3).

In our study, we measured the conversion bias among conversion tracts for which the two parental haplotypes present AT_f/GC_f-polymorphisms, i.e. only a subset of all observed conversion tracts (see Main Text). As we assume that there is no correlation between the nature (GC or AT) of neighboring alleles on a same haplotype, the nature of N::N mismatches present in chromatid I heteroduplex is independent of the identity of mismatches in chromatid II heteroduplex. Hence, we can assume that in Table S3, each N::N parental mismatch has a probability r to be W::S or S::W and $1-r$ to be W::W or S::S. Thus, we can derive the relative abundance of GC_f and AT_f haplotypes expected among the subset of conversion tracts with AT_f/GC_f-polymorphism, for each of the five cases (Table S4). From this, we can compute the expected relative abundance of GC_f and AT_f haplotypes, among the subset of conversion tracts with AT_f/GC_f-polymorphism:

$$\begin{aligned} n_{GCf} &= r^2(2 - p_W - p_S)x + rp_Sx + p_S(1 - x) \\ n_{ATf} &= r^2(2 - p_W - p_S)x + rp_Wx + p_W(1 - x) \end{aligned} \quad (5)$$

Thus,

$$n_{GCf} - n_{ATf} = (p_S - p_W)(1 + x(r - 1)) \quad (6)$$

According to our model $p_S > p_W$. Given that x and r are probabilities, $(1 + x(r - 1))$ is positive. Hence, our model predicts that $n_{GCf} - n_{ATf}$ is positive, in agreement with the observation that conversion tracts with AT_f/GC_f-polymorphism lead to an over-transmission of GC_f haplotypes (Table 2).

Let us now consider the average length of GC_f and AT_f conversion tracts: L_{GCf} and L_{ATf} . Using the relative abundance of the different tracts and their length (Table S4), we can derive the following results:

$$\begin{aligned} L_{GCf} &= \frac{r^2(2 - p_W - p_S)xL_1 + rp_SxL_2 + p_S(1 - x)L_1}{n_{GCf}} \\ L_{ATf} &= \frac{r^2(2 - p_W - p_S)xL_1 + rp_WxL_2 + p_W(1 - x)L_1}{n_{ATf}} \end{aligned} \quad (7)$$

Thus

$$L_{GCf} - L_{ATf} = \frac{r^3 x^2 (2 - p_w - p_s)(p_s - p_w)(L_2 - L_1)}{n_{GCf} n_{ATf}} \quad (8)$$

Given that $L_2 > L_1$, and $p_s > p_w$ this value is expected to be positive. Hence, in agreement with our observations (see Main Text), our model predicts that, on average, GC_f conversion tracts should be longer than AT_f conversion tracts.